Check for
updates

Examining the effectiveness of brief interventions to strengthen a positive implicit relation between women and STEM across two timepoints

Lynn Farrell[1,2] · Finiki Nearchou[1] · Louise McHugh[1]

Abstract

The current study assessed the impact of three brief interventions aimed at influencing implicit gender bias in Science, Technology, Engineering and Math (STEM) fields. The reduction of this bias is a key consideration as it remains a major barrier to gender equality in STEM. The interventions (psychoeducation, exposure to positive counter-stereotypical exemplars and perspective-taking) were compared to a control group at two timepoints. Gender-STEM bias was assessed at the implicit level (via the Implicit Relational Assessment Procedure; IRAP) and the explicit level (via rating scales). Two hundred and ten adults (58.1% women) completed one of the four study conditions. Results indicated that implicit gender-STEM bias is malleable, at least in the short term. At Time 1, intervention groups showed higher levels of implicit pro-Women-STEM bias and lower levels of pro-Men-STEM bias than the Control group. Psychoeducation appeared most effective, followed by exposure to positive counter-stereotypical exemplars. Results from Time 2 presented a more complex picture of implicit bias change, as the control group exhibited an increased pro-Women-STEM bias that was unexpected. The implications of these findings are discussed in relation to refining the interventions, the malleability of gender-STEM bias and the measurement of implicit bias across contexts.

Keywords Implicit bias · Implicit relational assessment procedure · Gender bias · STEM · Intervention

✉ Lynn Farrell
L.Farrell@qub.ac.uk

Extended author information available on the last page of the article

1 Introduction

The underrepresentation of women in Science, Technology, Engineering and Math (STEM) fields continues to be a global concern (Sánchez de Madariaga et al. 2012; Schwab et al. 2016; UNESCO 2017). Research has identified gender stereotypes along with gender bias (often implicit) to be one of the main barriers for women in STEM (National Academy of Sciences 2006; see Carnes et al. 2012; Master and Meltzoff 2016; Sánchez de Madariaga et al. 2012). Implicit bias associating men with STEM and women with the Arts has been found across at least 66 nations (Miller et al. 2015). Master and Meltzoff (2016) suggest that two main gendered stereotypes are involved in men-STEM bias: (a) the belief that men are a better 'fit' for STEM subjects, with scientists often perceived as more similar to Western masculine stereotypes (e.g., Carli et al. 2016; Carnes et al. 2015; Gatta and Trigg 2001; Kaatz and Carnes 2014), and (b) the belief that men have more 'ability' for STEM subjects than women, with more potential for achievement in these fields (e.g., Gatta and Trigg 2001; Leslie et al. 2015; Margolis et al. 2000; Schmader et al. 2004; Spencer et al. 1999; Summers 2005). These stereotypes have been found to negatively influence STEM interest (e.g., Ehrlinger et al. 2017; Master et al. 2016), performance (Kiefer and Sekaquaptewa 2007; Smeding 2012) and sense of belonging (e.g., Cheryan et al. 2009) among women and girls. A bias in favor of men has also been found in STEM-related performance evaluations (e.g., Axelson et al. 2010; Isaac et al. 2011), letters of recommendation (e.g., Schmader et al. 2007), academic publications (e.g., Knobloch-Westerwick et al. 2013), and hiring decisions (e.g., Moss-Racusin et al. 2012; Reuben et al. 2014; though cf. Williams and Ceci 2015). Reducing implicit gender-STEM bias has the potential to positively impact on the recruitment and retention of women in STEM subjects and careers by making women succeeding in STEM as normative as men.

1.1 Malleability of implicit bias

Recent research suggests that implicit bias is malleable and can be influenced, at least in the short term, by a number of contextual factors (e.g., Rudman et al. 2001; Dasgupta and Asgari 2004; Blair et al. 2001; Todd et al. 2011). There is less research on the malleability of implicit gender stereotypes (see Lenton et al. 2009 for a meta-analysis), and fewer still focused on implicit gender-STEM bias change post-intervention (see Jackson et al. 2014). Additionally, very few studies have assessed implicit bias change at more than one timepoint (see Lai et al. 2013), which makes hypothesizing about long-term implicit bias change difficult. The current study, influenced by the extant literature, will add to the research in this area by examining three brief gender-STEM bias interventions at two timepoints. The interventions consist of psychoeducation, exposure to counter-stereotypical exemplars and perspective-taking.

1.1.1 Psychoeducation

Psychoeducation is arguably one of the most popular approaches aimed at reducing bias including formats such as diversity training (e.g., Jackson et al. 2014), bias literacy workshops (e.g., Carnes et al. 2012) and video-based interventions (e.g., Moss-Racusin et al. 2018). It typically involves lessening individuals' reliance on stereotyping and better informing them of the nature of bias itself in order to both normalize and counter such modes of thinking. The effectiveness of psychoeducation interventions has rarely been evaluated however, particularly at the implicit level, and even less so specific to gender-STEM bias (Jackson et al. 2014; see Moss-Racusin et al. 2018). Psychoeducation has influenced outcomes such as awareness of bias and self-efficacy to address gender bias issues, even at follow up assessments (Carnes et al. 2015; Moss-Racusin et al. 2016). However, its influence on implicit gender bias has been mixed (Carnes et al. 2015; Moss-Racusin et al. 2016, 2018) and it is unclear how implicit gender-STEM bias assessed via more specific measures such as the Implicit Association Test (IAT; Greenwald et al. 1998) may have been influenced (see Moss-Racusin et al. 2018). Jackson et al. (2014) assessed implicit attitudes towards women in STEM following brief diversity training using a paper-based, personalized version of the Go/No-Go Association Task. They found positive implicit attitudes towards women in STEM increased post-intervention compared to a control group but only among men. Explicit attitudes were positive pre-intervention and did not significantly increase post-intervention. This research suggests that psychoeducation may be an effective tool when empirically led and evaluated (see Moss-Racusin et al. 2018).

The current study's psychoeducation involved an individual reading piece to examine whether this information alone would influence implicit gender-STEM bias. It included brief descriptions of implicit bias and approaches to overcome bias, non-confrontational, inclusive language (e.g., acknowledging that we all hold biases) and evidence-based content such as information disconfirming stereotypes and highlighting gender disparities in STEM. This information was influenced by the workshops of Jackson et al. (2014) and Carnes et al. (2012) and featured some of the components Moss-Racusin et al. (2014) recommend for diversity interventions.

1.1.2 Exemplars

Counter-stereotypical exemplars have previously influenced implicit ageist bias (Cullen et al. 2009), racial attitudes (Dasgupta and Greenwald 2001) and gender stereotypes (e.g., Dasgupta and Asgari 2004). Exposure to counter-stereotypical scientist exemplars allows individuals to update their beliefs regarding who may succeed in STEM or what type of person is the norm in STEM (often considered to be white and male), potentially highlighting heterogenous relations that may be present even within stereotypes (see Lenton et al. 2009). Specific to gender-STEM bias, exposure to exemplars has produced mixed results. For example, Ramsey et al. (2013; Study 2) found positive exemplars of women in STEM and men in the humanities increased women's implicit STEM identification but implicit gender stereotypes were not significantly reduced (though see also Stout et al. 2011; Study 2). Explicit

identity and stereotypes remained unchanged also. Joy-Gaba and Nosek (2010) exposed participants to 8 admired female scientists but gender-academic stereotypes were not significantly reduced. However, if women were already positively associated with STEM (albeit less so than men) then a post-intervention increase in this association may have been too small to be detected, particularly on a relative measure such as the IAT (see Joy-Gaba and Nosek 2010).

There is a suggestion it may be necessary to include a negative contrast category (e.g., disliked male scientists) and make the variable of interest (e.g., gender) salient in order to influence implicit bias. However, this would have implications for the applicability of this intervention as it would be undesirable to lower the evaluation of one social group to increase that of another in pursuit of gender equality (Joy-Gaba and Nosek 2010). A negative contrast category (i.e., disliked male scientists) was not included in the current study's exemplar-based intervention. The exemplars made gender salient by highlighting that these were female scientists who may be lesser known as their achievements were initially overlooked or underappreciated.

1.1.3 Perspective-taking

Perspective-taking involves adopting the perspective of another person by considering their psychological experiences (Todd et al. 2011) or 'seeing' the world through their eyes. Perspective-taking may increase empathy and prosocial behavior towards the target group (e.g., Shih et al. 2009). It can increase the overlap between the self and the 'other'. This appears to diminish the accessibility of stereotypes and increase positive evaluations of the perspective-taking target which may then generalize to the target's social group. This effect may occur unconsciously (Galinsky and Moskowitz 2000).

To date, this approach has been used to reduce ageist (e.g., Galinsky and Moskowitz 2000; Yee and Bailenson 2006) and implicit racial bias (e.g., Todd et al. 2011; Todd et al. 2012). It has yet to be tested on implicit gender-STEM bias specifically. As bias towards women in STEM appears to be influenced by stereotypes about both gender and STEM, the narrative perspective-taking task (e.g., write about a day in the life of a female scientist) was adopted for the current study as it has previously reduced stereotyping, potentially targeting more cognitive aspects (e.g., Galinsky and Moskowitz 2000; see Dovidio et al. 2004; Todd et al. 2011).

1.2 Implicit Relational Assessment Procedure

Most research into implicit gender-STEM bias has relied upon relative measures of implicit responding such as the IAT. These measures are limited in the level of detail they can provide, usually comparing STEM subjects relative to Arts subjects, and unable to separate out the components of the bias. For example, a change in an IAT score assessing gender-STEM bias could reflect a change in bias towards men, women, or both (see Lai et al. 2014). When assessing the effectiveness of bias interventions, a measure of implicit bias that can analytically separate the components of a particular bias (e.g., examine Men-STEM and Women-STEM relations) may,

therefore, provide an advantage by more clearly explicating the relations influenced by the intervention. The current study, therefore, utilized the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al. 2006) as a measure of implicit bias that can provide an individual analysis of each of the assessed relations, which has previously been employed in the domain of gender-STEM bias (Farrell et al. 2015; Farrell and McHugh 2017).

The IRAP developed as a measure from Relational Frame Theory (RFT; Hayes et al. 2001), a contemporary account of human language and cognition based on behavior-analytical principles. According to RFT, the key behavioral process underlying language and cognition is derived relational responding (see Barnes-Holmes et al. 2004 for a more detailed explanation). Both implicit and explicit biases are forms of relational responding that vary in terms of levels of derivation (i.e., how practiced they are) and complexity (Hughes et al. 2012). Implicit responses represent an individual's initial response to stimuli that are relatively lower in terms of relational complexity and derivation (i.e., highly practiced responses). Explicit responses, on the other hand, are more elaborated and can involve relatively higher levels of complexity and derivation (i.e., they can be more novel; Hughes et al. 2012).

The IRAP requires participants to confirm or deny a particular belief by responding to the relation between a label statement (e.g., 'Women more suited to', 'Men more suited to') and a target word (e.g., 'Science', 'Arts') using the given response options (e.g., 'True' or 'False') across a number of blocks of trials. The combination of each label statement category and each target word category results in four overarching IRAP trial-types for which response biases may be assessed (e.g., Men-STEM, Men-Arts, Women-STEM, Women-Arts). Correct responding during the IRAP is governed by pre-determined rules which switch between the blocks. Half of the blocks require stereotype-consistent responding (e.g., Men more suited to STEM and women more suited to Arts) while the other half require stereotype-inconsistent responding (e.g., Women more suited to STEM and men more suited to Arts). The IRAP assesses how fluently (i.e., quickly and accurately) participants are able to relate stimuli on the basis of these pre-determined rules under time pressure. The basic premise is that if a participant's initial brief relational response matches the response required by the current IRAP rule (e.g., responding 'True' to 'Men more suited to Physics') then responding will be quicker than when participants must respond against their initial relational response (e.g., responding 'False' to 'Men more suited to Physics'). The response latency differential between stereotype-consistent and -inconsistent blocks is used to infer a person's implicit bias towards the stimuli influenced by historical and current contextual variables (Barnes-Holmes et al. 2010). The IRAP is non-relative insofar as the fact that each of its four trial-type scores are calculated independently from the other trial-types (see Hughes et al. 2017; see Hussey et al. 2016). That is, for the trial-type relating men with STEM subjects, participants' response times are compared when they responded with 'True' versus 'False' only and are not compared with their responses to 'Men more suited to' Arts subjects.

A recent IRAP study (Farrell and McHugh 2017) on implicit gender-STEM bias among STEM and non-STEM students demonstrated the finer grained analysis

provided by the IRAP. By examining the components of this bias separately, Farrell and McHugh (2017) found that in addition to exhibiting an implicit pro-Men-STEM bias (demonstrated by the other STEM and non-STEM groups), women in STEM also had a significant pro-Women-STEM bias. This may contribute to the lower Male-STEM/Female-Arts bias that women in STEM exhibit on the IAT (Nosek and Smyth 2011; Smeding 2012; Smyth and Nosek 2015). Farrell and McHugh (2017) also found that the other groups (non-STEM students and men in STEM) directionally exhibited implicit pro-Women-STEM responding, though it was very weak and non-significant. On the basis of current stereotypes, one may have expected these groups to have an anti-Women-STEM bias instead. These findings have implications for intervention research as this weak implicit positive relation between women and STEM may be amenable to influence. Rather than lessen the belief that men are suited to STEM, the aim is to strengthen a positive relation between women and STEM to make women's suitability for STEM as normative as men's. This may have a beneficial impact upon behavior and attitudes towards women in STEM and challenge implicit bias in this domain (Farrell and McHugh 2017).

1.3 The current study

The current study aimed to influence implicit gender-STEM bias through brief interventions. At the time of writing no published IRAP study had examined the malleability of implicit gender bias in STEM. We hypothesized that an effective implicit gender-STEM bias intervention would strengthen a pro-Women-STEM relation among participants. It was less clear how it would affect pro-Men-STEM bias, though we did not anticipate a reversal in this bias (i.e., anti-Men-STEM bias). We focused on analyzing the two STEM-focused trial-types of the IRAP (i.e., Men-STEM and Women-STEM) as they were the targets of our intervention and seem to elicit stronger relational responding, driving differential levels of gender-STEM bias (see Farrell et al. 2015; Farrell and McHugh 2017).[1]

This study addresses limitations within the literature by: (a) using the IRAP as a measure of implicit bias to probe more specific relations between STEM and gender post-intervention; (b) assessing the impact of the interventions both implicitly and explicitly; (c) assessing levels of implicit and explicit gender-STEM bias at two timepoints—immediately post-intervention and on the following day, and (d) comparing three interventions against a control group to determine if one is more effective than the others. For example, will increasing empathy and the self-other overlap via perspective-taking decrease the stereotypes that drive implicit gender-STEM bias (i.e., increasing relations of similarity or co-ordination and decreasing opposition relations between groups; Edwards et al. 2017)? Or is it more effective to target the relation between women and STEM more directly by highlighting positive exemplars or raising awareness about implicit bias against women in STEM (i.e.,

---

[1] Details of the Arts-focused trial-types' analysis can be made available on request but no significant intervention effect was detected on these trials at either timepoint.

allowing further opportunities to derive positive relations of co-ordination between women, competency and STEM, and reduce relations of opposition)?

There is a suggestion that implicit bias is only malleable in the short-term (Lai et al. 2014) and it is unlikely that a single intervention exposure would generate lasting change (Lenton et al. 2009). However, one may surmise that the more opportunities an individual has to practice or experience counter-stereotypical relations the better. The interventions included in the current study, therefore, aim to highlight contextual factors which may influence gender-STEM bias should they be incorporated into the environment (e.g., more positive, competent examples of female scientists in the media). It is important to explore these factors so as to intervene against gender-STEM bias within our culture.

2 Method

Ethical approval was obtained from the Human Research Ethics Committee—Humanities at University College Dublin prior to beginning data collection with informed consent obtained from all participants.

2.1 Sample size

Using G*Power 3.1 a minimum sample size of 168 (42 participants per group) was determined based on the following parameters: a moderate effect size of .15(f), alpha of .05, power of .8, 8 groups (4 interventions conditions by 2 genders), 2 measurement points, and correlations between variables of .5. Given the exploratory nature of the study the decision was made to collect as many participants as possible within a pre-allocated data collection period from October 2016 until December 2017, with the stipulation that each group would have at least 42 participants. This was adequate for the IRAP analyses also (see Vahey et al. 2015). Our rationale for collecting more participants was based upon attrition concerns. IRAP attrition with an adult population can range from 5% (e.g., Farrell and McHugh 2017) to 52% (Hooper et al. 2010). Therefore, a higher target sample size was deemed necessary to ensure the minimum sample size was achieved.

2.2 Participants

A volunteer, convenience sampling method was used to recruit participants, along with an online recruitment pool moderated by the host institute. Adults were the focus of the current study to build on previous IRAP work examining implicit gender-STEM bias (e.g., Farrell et al. 2015). Adults can influence important educational decisions that students make such as subject choice and career path (e.g., influence of parents and teachers; Akosah-Twumasi et al. 2018). Their support can also influence students' persistence and engagement in STEM fields, promoting a greater sense of belonging within these fields, for example (e.g., influence of family and friends; Rosenthal et al. 2011). However, implicit gender-STEM bias has been

found among both student and non-student adult groups (e.g., Farrell et al. 2015; Farrell and McHugh 2017; Nosek et al. 2009). Therefore, it is important to determine whether adult implicit gender-STEM biases can be reduced.

Inclusion criteria were fluent English, age 18+ and normal or corrected to normal vision. Exclusion criteria were women who identified as scientists or STEM students, incomplete data through failure to return for session 2 and/or failure to maintain performance criteria for the IRAP during either session. Women studying/working in STEM were excluded as previous research demonstrated that they already possessed a significant pro-Women-STEM bias (Farrell and McHugh 2017), which was the main target of the current interventions. We did not wish to risk possible ceiling effects and/or artificially inflating the results of any of the intervention conditions by including this group in this initial study of gender-STEM malleability.

By December 2017, 227 eligible participants had participated in the study. Five participants did not return to complete the second session of the study and so were excluded from the analysis. Eleven failed to maintain the performance criteria for the IRAP (see Sect. 3.1 below) either during session 1 or session 2 and were excluded. Finally, one participant in the Perspective-Taking group failed to follow the task instructions accurately (i.e., failed to write a first-person narrative) and so their data were excluded. This resulted in a final sample of 210 participants (Control $N=53$; Exemplar $N=52$; Psychoeducation $N=52$ and Perspective-Taking $N=53$). The majority of the sample were students (both STEM and non-STEM; 90%). The remaining sample consisted of non-students that were either employed (7.6%) or unemployed (2.4%). Participants mainly identified as Irish (61.9%), with the remainder a diverse mix of nationalities such as American (10.95%), Indian (6.2%), Chinese (2.4%), and Italian (1.9%). The mean age of participants was 23.4 ($SD=7.6$), with a range of 18-59 years. In accordance with recommendations (e.g., Ansara and Hegarty 2014), participants self-identified their gender. The majority were women (58.1%), with the remainder identifying as men (41.4%) or non-binary (.5%).

2.3 Materials

2.3.1 Intervention materials

Each participant completed one of four intervention conditions—Control, Exemplars, Psychoeducation or Perspective-Taking. All intervention materials were paper based. The Control condition involved ten color pictures of animals. Beneath each picture was the animal's name followed by a paragraph of similar length (170–189 words) discussing, for example, their eating habits. This was followed by a ten-question multiple choice quiz (i.e., choose a, b, or c) based on the animals featured.

The Exemplar condition included eleven black and white photographs of female scientists. The female scientists were described thusly: "You will be presented with a number of black and white photographs of female scientists who accomplished great work in their field. While they overcame obstacles to achieve this work, they were either overlooked or under-appreciated at the time of their initial success. For this reason, some of these scientists may not be known to you". Above each photograph

was the scientist's name and below was a paragraph (116–153 words) discussing their achievement, field, and, where applicable, how they were overlooked (e.g., not included in a Nobel Prize awarded to their colleagues for shared work). This was followed by a ten-question quiz (multiple choice bar one question).

The Psychoeducation condition involved information relevant to gender-STEM bias and was described thusly: "The following information is concerned with the issue of gender bias towards women in Science, Technology, Engineering and Math (STEM). It aims to show how the stereotypes that contribute to this bias can be incorrect and briefly details strategies to overcome this bias." Citations were provided where relevant to demonstrate that the information was supported by research. Gender-STEM stereotypes and implicit bias were discussed. Stereotype disconfirming information was included (e.g., it was noted that there is no conclusive evidence that men are biologically better at STEM subjects) and strategies to overcome bias were suggested (e.g., imagining counter stereotypes and increasing awareness of implicit bias). Questions were interspersed throughout the Psychoeducation information to engage the reader such as "If I were to ask you to draw a scientist, what would you draw?" This piece consisted of 2457 words. It was followed by a ten-question quiz with a mixture of multiple choice, True/False and open-ended questions.

Finally, the Perspective-Taking condition consisted of a color photograph of a female scientist at work alone in a lab-setting. Both above and beneath were the task instructions. Participants were instructed to: "Imagine a day in the life of this individual as if you were that person and not yourself, looking at the world through her eyes and walking through the world in her shoes. What are your thoughts, feelings, and experiences? What scientific project are you working on? As you write remember that her perspective is your perspective as you are her." Participants had a maximum of 10 min to complete this task—5 min maximum to examine the photograph and 'gather their thoughts' and 5 min thereafter to write their narrative. Narratives had to be written in the first-person.

2.3.2 IRAP

The IRAP is a computer-based response latency task. Stimuli included two label statements ("Women more suited to" or "Men more suited to"), twelve target words and two response options ("True" and "False"). The target words consisted of six STEM subjects (Science, Maths, Physics, Engineering, Computing and Chemistry) and six Arts subjects (Arts, English, Drama, French, Music and History). The combination of each of the label statement categories with each target word category resulted in four IRAP trial-types—Men-STEM, Men-Arts, Women-STEM and Women-Arts. Participants were presented with a label statement at the top of the screen (e.g., "Women more suited to") and one of the twelve target words in the center of the screen (e.g., "Maths"). The response options "True" and "False" were displayed at the bottom left and right of the screen. To respond with "True" participants were required to press the 'd' key on the computer keyboard while a response of "False" required pressing the 'k' key. Response options remained fixed

throughout. Each label statement appeared once with each target word in quasi-random order resulting in 24 trials in each block of the IRAP.

Correct responding during each IRAP block was governed by a rule which was presented to participants before each block and alternated between blocks. In certain blocks the rule required stereotype-consistent responding (i.e., Please respond as if men are more suited to Science subjects and women are more suited to Arts subjects) while for the remainder of the blocks the rule switched, requiring stereotype-inconsistent responding (i.e., Please respond as if women are more suited to Science subjects and men are more suited to Arts subjects). The block order (i.e., consistent or inconsistent block first) was counterbalanced across participants. If a participant made an incorrect response a red 'X' would appear onscreen until a correct response was subsequently made. A red exclamation mark would appear at the bottom of the screen to warn participants when the response time limit was near (i.e., 2 s) though the stimuli remained onscreen until a correct response was made. At the end of each block participants were presented onscreen with their mean accuracy and median response latency for that block.

An odd–even split-half reliability procedure was used to calculate the reliability of the IRAP scores (see De Houwer and De Bruycker 2007) using the Spearman-Brown formula. All 4 IRAP trial-types were included in line with previous reliability assessments. At Time 1, split-half reliability was .59 for the IRAP scores. At Time 2, it was .58. These values are reasonably strong and in line with previous IRAP studies (e.g., Farrell and McHugh 2017; McKenna et al. 2016; Remue et al. 2013). These results also compare well to other latency-based measures of implicit bias (Golijani-Moghaddam et al. 2013). Test–retest reliability of the current IRAP showed a moderate significant correlation between the timepoints for each of the relations ($N = 210$; Men-STEM $r = .30$, $p < .001$; Women-STEM $r = .37$, $p < .001$).

2.3.3 Rating scales

Explicit bias was assessed using twelve rating scales via Qualtrics (2017)—one for each of the STEM and Arts subjects used in the IRAP (see Farrell and McHugh 2017). Participants rated on an 11-point scale whether males or females were more suited to each subject. A score of 6 indicated that both males and females were rated as equally suitable for the subject. Scores above 6 indicated females were deemed more suitable while scores below 6 indicated males were. The STEM scales demonstrated high reliability (Time 1: $\alpha = .807$; Time 2: $\alpha = .823$) while the Arts scales were moderately reliable (Time 1: $\alpha = .699$; Time 2: $\alpha = .678$; Hinton et al. 2004).

2.4 Procedure

Participants completed two sessions with implicit bias (via the IRAP) and explicit bias (via the rating scales) measured at two timepoints—once immediately post-intervention and once the following day (a minimum of 16 h after session one). Both sessions took place in a quiet room to minimize distractions with the same female experimenter throughout. The study was initially described as being concerned with

assessing the impact of information processing on rule-governed responding so as to reduce the risk of a potential sampling bias, whereby only those interested in gender bias would volunteer. It also reduced the potential for participants' responses to be influenced by knowledge of the study's aim.

2.4.1 Session 1

Participants were asked to select a number between 1 and 4 in order to randomly decide which condition they would complete. They were unaware which number corresponded to which condition or the topics covered. Participation levels for each condition were monitored throughout to ensure there were at least 42 participants in each. Once they had selected a number, participants completed the corresponding condition—Control, Exemplar, Psychoeducation or Perspective-Taking. The experimenter left the room while participants completed this task. Only the Perspective-Taking task had a time limit (10 min maximum) for completion. Generally, participants completed the other 3 intervention tasks within 25 min. When completing either the Control, Exemplar or Psychoeducation tasks, participants completed a quiz afterwards to ensure that they engaged with the material. After the quiz, the experimenter informed participants of the answers so as to reinforce the correct material. The final sample of participants' scores on these quizzes were in the range of 7–10 correct answers out of 10 questions. This was deemed acceptable as it was above chance level responding (5/10).

Participants then completed the IRAP. Task instructions and visual samples of the IRAP trials were given to participants. The key point conveyed was that correct responding was governed by the rule given to them and not their own opinion. The experimenter remained in the room with the participant until they passed the practice blocks. In order to progress to the test blocks, participants were required to achieve a mean accuracy of $\geq 80\%$ and median response latency of ≤ 2 s on two consecutive practice blocks. If a participant completed a maximum of 8 practice blocks without achieving the mastery criteria, then they were thanked for their participation and excused from the study. If the participant achieved the mastery criteria, they proceeded to the six test blocks, at which point the experimenter left the room. Short breaks between the blocks were advised to avoid any fatigue effects which could have led to response errors.

Once the participants completed the six test blocks the experimenter re-entered the room to present the rating scales. Finally, the demographic questionnaire was completed. Before leaving, participants were reminded to return the next day for their pre-arranged session 2 timeslot. In total, session 1 took approximately 45–50 min on average.

2.4.2 Session 2

Participants completed the same IRAP, followed by the rating scales and demographic questionnaire. Finally, participants were fully debriefed that the study was concerned with assessing the malleability of gender-STEM bias and were compensated €10 for completing both sessions of the study (€5 per session). No participant

Table 1 Means and standard deviations for implicit and explicit scores by intervention group. Means with superscripts differed significantly ($p < .025$)

Time	Score type	Control M (SD)	Exemplar M (SD)	Psychoeducation M (SD)	Perspective-taking M (SD)
1	Men-STEM IRAP	.43 (.37)	.24 (.40)	.17 (.42)[a]	.31 (.48)
2	Men-STEM IRAP	.48 (.33)	.39 (.40)**	.40 (.44)*	.31 (.48)
1	Women-STEM IRAP	− .11 (.42)	− .37 (.41)[a]	− .39 (.37)[a]	− .24 (.39)
2	Women-STEM IRAP	− .36 (.41)*	− .31 (.45)	− .31 (.36)	− .31 (.38)
1	STEM	32.58 (4.13)	33.71 (4.43)	34.90 (2.32)[a]	32.88 (4.10)
2	STEM	33.29 (3.83)	33.58 (3.97)	35.24 (1.81)[a]	33.46 (3.52)
1	Arts	38.10 (3.37)	37.96 (2.84)	37.08 (1.92)	38.36 (3.24)
2	Arts	37.83 (2.57)	37.44 (3.25)	36.56 (1.51)[a]	37.94 (2.98)

[a]Signifies a difference from the Control group at that time point. An Asterix indicates a significant difference in scores between Time 1 and Time 2. Two asterisks indicate a significant difference in scores between Time 1 and Time 2 for women only. Implicit scores $N = 210$; Explicit scores $N = 200$. Excluding non-binary participant resulted in slight changes to Perspective-Taking *Means*: Women-STEM T1 $M = −.23$ (.39) and T2 $M = −.30$ (.37). STEM T1 $M = 32.82$ (4.08); T2 $M = 33.41$ (3.54); Arts T1 $M = 38.41$ (3.26); T2 $M = 37.98$ (2.996)

withdrew from the study as a result of the deception involved. This session took approximately 15-20 min on average.

3 Results

3.1 IRAP

For inclusion in the subsequent analysis participants had to maintain a median latency of ≤ 2 s and accuracy of $\geq 80\%$ on average across the 3 consistent and 3 inconsistent blocks separately in the IRAP. Response latencies were the primary data, defined as the time from the onset of a trial to the first correct response. This data was transformed into *D*-IRAP scores (see Barnes-Holmes et al. 2010). At each timepoint participants produced a *D*-score for each of the 4 IRAP trial-types—Men-STEM, Men-Arts, Women-STEM and Women-Arts. As noted, we focused on the STEM IRAP trials in this analysis. First a Multivariate Analysis of Variance (MANOVA) was conducted for each intervention condition to ensure that *D*-IRAP scores at both timepoints were not significantly influenced by IRAP block order (consistent or inconsistent block first). There were no significant effects of block order (all p's $> .09$), therefore, it was dropped from subsequent analyses.

3.1.1 IRAP immediately post-intervention

We first consider *D*-IRAP scores at Time 1 to assess whether the interventions had an influence on gender-STEM bias immediately post-intervention. Table 1 presents

the mean scores for the Men-STEM and Women-STEM IRAP trial-types at Time 1 for each Intervention group. Scores in a positive direction indicated that men were related to STEM (pro-Men-STEM) and women were not related to STEM (anti-Women-STEM). While scores in a negative direction indicated women were related to STEM (pro-Women-STEM) and men were not related to STEM (anti-Men-STEM). Larger scores (in terms of the absolute number) indicated a larger response bias.

From Table 1 we can see that all groups descriptively exhibited pro-Men-STEM and pro-Women-STEM response biases to various degrees. The Control group had the highest mean level of pro-Men-STEM response bias, with a weak positive Women-STEM relation similar to previous research (see Farrell and McHugh 2017). The intervention groups showed relatively higher levels of pro-Women-STEM bias and lower levels of pro-Men-STEM bias, particularly the Psychoeducation group.

A 2-way MANOVA with intervention condition (Control, Psychoeducation, Exemplar, Perspective-Taking) and gender (men or women)[2] as independent variables was conducted on the Men-STEM and Women-STEM IRAP scores. Gender and intervention did not significantly interact, $p = .334$, $\eta_p^2 = .02$, nor was there a significant gender effect, $p = .171$, $\eta_p^2 = .02$. There was, however, a statistically significant effect of intervention condition, Wilks Lambda $= .89$, $F(6, 400) = 3.88$, $p = .001$, $\eta_p^2 = .06$. There was a statistically significant main effect of intervention for the Men-STEM, $F(3, 201) = 3.94$, $p = .009$, $\eta_p^2 = .06$, and the Women-STEM IRAP scores, $F(3, 201) = 4.99$, $p = .002$, $\eta_p^2 = .07$. Scheffe post hoc tests with Bonferroni corrected alpha levels of .025[3] revealed that for the Men-STEM IRAP scores, only the Psychoeducation group was significantly different from the Control group ($p = .018$). The Men-STEM scores of the Control group were .26 higher than the Psychoeducation group, 95% CI [.03, .49], indicating a lower level of pro-Men-STEM bias among the Psychoeducation group immediately post-intervention. Scheffe post hoc tests also revealed that for Women-STEM IRAP scores both the Exemplar ($p = .013$) and the Psychoeducation ($p = .006$) groups were significantly different from the Control group. There was a difference between the Women-STEM scores of the Control group and the Exemplar and Psychoeducation groups of .26, 95% CI [.04, .48] and .28, 95% CI [.06, .50] respectively. This indicated a higher level of pro-Women-STEM bias among the Exemplar and Psychoeducation groups immediately post-intervention.

---

[2] As only a single participant identified as non-binary, only men and women were included in this analysis. The same approach was taken for the subsequent Time 2 and explicit analyses (i.e., non-binary participant included in descriptive statistics in Table 1 but excluded from subsequent MANOVAs and post hoc tests).

[3] Bonferroni corrected alpha levels were applied where appropriate throughout, dividing the alpha level by the number of scores (.05/2; α = .025). This correction accounted for multiple tests being conducted but was not as conservative as a more stringent application of Bonferroni which could reduce statistical power and increase the possibility of Type II error (e.g., Nakagawa 2004; Perneger 1998; Rothman 1990).

3.1.2 IRAP administered the next day

We next considered the IRAP results from session 2 (the following day). Table 1 presents the mean scores for the Men-STEM and Women-STEM IRAP trials at Time 2 for each intervention group. The Control group still showed a pro-Men-STEM bias. Surprisingly, they exhibited a higher mean pro-Women-STEM bias than they had at Time 1. The pro-Men-STEM biases of the Exemplar and Psychoeducation groups had also increased at Time 2 while their pro-Women-STEM biases had reduced slightly. The Perspective-Taking group showed a similar level of pro-Men-STEM bias at Time 2 while their pro-Women-STEM bias had increased slightly.

The data were explored using a repeated-measures, mixed ANOVA with gender and intervention condition as the between-participants variables, and time (Time 1: Immediately post-intervention, and Time 2: The following day) and IRAP trial-type (Men-STEM and Women-STEM) as the within-participants variables. IRAP scores at Time 1 and Time 2 were the dependent variables. There was a main effect for IRAP trial-type, $F(1, 201) = 378.99$, $p < .001$, $\eta_p^2 = .65$. There were two significant two-way interactions—one between trial-type and time, $F(1, 201) = 9.14$, $p = .003$, $\eta_p^2 = .04$ and the other between time and intervention, $F(3, 201) = 5.83$, $p = .001$, $\eta_p^2 = .08$. Both of these were moderated by the significant 4-way interaction between IRAP trial-type, time, gender and intervention group, $F(3, 201) = 2.79$, $p = .042$, $\eta_p^2 = .04$. Due to the presence of this 4-way interaction we ran separate repeated-measures, mixed ANOVAs on each intervention group (the focal variable) with gender as the between-participants variable and time and trial-type as the within-participants variables.

There was a main effect of trial-type for the Control group, $F(1, 51) = 128.24$, $p < .001$, $\eta_p^2 = .72$. Additionally, there was an interaction between trial-type and time, $F(1, 51) = 14.01$, $p < .001$, $\eta_p^2 = .22$. Follow up Bonferroni corrected paired sample t-tests found a significant difference between Time 1 and Time 2 for the Women-STEM IRAP trial, $t(52) = 3.83$, $p < .001$, Cohen's $d = .53$, 95% CI [.12, .37]. Control participants exhibited a stronger pro-Women-STEM response bias at Time 2.

The Exemplar group showed a main effect for trial-type, $F(1, 50) = 90.00$, $p < .001$, $\eta_p^2 = .64$ and a significant interaction between time and gender, $F(1, 50) = 4.38$, $p = .041$, $\eta_p^2 = .08$. In order to further examine this, repeated-measure ANOVAs were conducted on men and women from the Exemplar group separately. Both showed a main effect of trial-type, Men: $F(1, 22) = 46.98$, $p < .001$, $\eta_p^2 = .68$; Women: $F(1, 28) = 43.32$, $p < .001$, $\eta_p^2 = .61$. Women also showed a main effect of time, $F(1, 28) = 12.24$, $p = .002$, $\eta_p^2 = .30$. Bonferroni corrected paired sample t-tests showed that this difference between Time 1 and 2 was statistically significant on the Men-STEM trial only, $t(28) = -2.75$, $p = .010$, Cohen's $d = .5$, 95% CI $[-.37, -.05]$. Pro-Men-STEM bias was significantly higher at Time 2 for women in the Exemplar group.

The Psychoeducation group exhibited a main effect of trial-type, $F(1, 50) = 99.84$, $p < .001$, $\eta_p^2 = .67$ and a main effect of time, $F(1, 50) = 13.59$, $p = .001$, $\eta_p^2 = .21$. Time and trial-type also interacted, $F(1, 50) = 4.32$, $p = .043$, $\eta_p^2 = .08$. Follow up Bonferroni corrected paired sample t-tests revealed a significant difference between Time 1 and Time 2 for the Men-STEM IRAP trial only, $t(51) = -3.90$, $p < .001$, Cohen's

$d = .54$, 95% CI $[-.35, -.11]$. Psychoeducation participants exhibited a stronger mean pro-Men-STEM bias at Time 2.

There was a main effect of trial-type for the Perspective-Taking group also, $F(1, 50) = 70.40$, $p < .001$, $\eta_p^2 = .59$, along with a significant interaction between trial-type, time and gender $F(1, 50) = 4.41$, $p = .041$, $\eta_p^2 = .08$. However, when following up this interaction for each gender separately using repeated-measures ANOVAs, there were no significant main or interaction effects (all p's > .1) bar the main effects of trial-type, Men: $F(1, 16) = 27.11$, $p < .001$, $\eta_p^2 = .63$; Women: $F(1, 34) = 47.19$, $p < .001$, $\eta_p^2 = .58$.

3.2 Rating scales[4]

3.2.1 Rating scales immediately post-intervention

A total score was calculated for the 6 Arts and 6 STEM subjects separately to produce one score for explicit STEM bias and one score for explicit Arts bias per participant. As each rating scale was an 11-point scale, this resulted in a maximum total score of 66 and a neutral sore of 36 (as 6 was designated as the neutral score within each scale). The mean explicit scores for each intervention group at Time 1 can be seen in Table 1. Overall, each group rated males as relatively more suited to STEM subjects (scores < 36) and females as relatively more suited to Arts subjects (scores > 36). The scores were generally not extreme, however. The Psychoeducation group in particular was close to the neutral point of 36 for both subjects.

A two-way MANOVA was conducted with two independent variables—gender and intervention—and two dependent variables—explicit STEM scores and explicit Arts scores at Time 1. The assumption of homogeneity of covariance matrices was violated, as assessed by Box's M test ($p < .001$). However, as the ratio of the largest to smallest sample size within the cells of the design was approximately 2:1, the current MANOVA was considered robust to this particular violation (Huberty and Olejnik 2006; Tabachnick and Fidell 2013). As a result, however, Pillai's trace is reported as the multivariate test statistic (Tabachnick and Fidell 2013).

The interaction effect between gender and intervention condition on the combined dependent variables was not statistically significant, $p = .213$, $\eta_p^2 = .02$. The main effect of intervention was significant however, Pillai's trace $= .07$, $F(6, 382) = 2.30$, $p = .034$, $\eta_p^2 = .04$; as was the main effect of gender, Pillai's trace $= .04$, $F(2, 190) = 3.92$, $p = .021$, $\eta_p^2 = .04$. Levene's test of equality of variances indicated violation of the assumption of homogeneity of variances (STEM $p < .001$, Arts $p = .001$; though cf. Zimmerman 2004), therefore, for follow up univariate ANOVAs

---

[4] There were ten multivariate outliers (4.8% of the sample) among the explicit data according to Mahalanobis distance (Tabachnick and Fidell 2007). These were removed from the explicit analyses only leaving **N** = 200 (199 without the non-binary participant). Six participants were flagged as extreme outliers. However, the 5% trimmed mean suggested that their scores did not greatly influence the explicit score means. Therefore, in line with Goodwin and Leech (2006), we retained these participants.

Welch's F and Games-Howell post hoc tests are reported, all with Bonferroni corrected alpha levels of .025.

There was a statistically significant main effect of intervention for the explicit STEM scores at Time 1, Welch's $F(3, 103.40) = 5.93$, $p = .001$, est. $\omega^2 = .07$, but not for the explicit Arts scores, $p = .043$, est. $\omega^2 = .03$. Games-Howell post hoc tests indicated that the Psychoeducation group's explicit STEM scores were significantly different from the Control group ($p = .004$). The Psychoeducation group was 2.3 higher than the Control group, 95% CI [.59, 4.05]. This indicated weaker Male-STEM explicit bias among participants in the Psychoeducation group immediately post-intervention as their scores were closer to the neutral point of 36. There was also a statistically significant main effect of gender for the explicit Arts scores at Time 1, Welch's $F(1, 145.94) = 5.34$, $p = .022$, est. $\omega^2 = .02$. Men had a higher explicit Female-Arts bias ($M = 38.47$; $SD = 3.33$) than women ($M = 37.47$; $SD = 2.53$).

3.2.2 Rating scales administered the next day

The mean total scores for each intervention group at Time 2 can be seen in Table 1. Again, there was a general Male-STEM and Female-Arts explicit bias, but scores were not extreme, particularly for the Psychoeducation group. A repeated-measures, mixed ANOVA was conducted with gender and intervention as the independent variables and time and explicit score type (STEM or Arts) as within-participants variables. Pillai's trace is reported as the multivariate test statistic (Tabachnick and Fidell 2013). There was a main effect of explicit score type, Pillai's trace = .36, $F(1, 191) = 108.45$, $p < .001$, $\eta_p^2 = .36$. There was a significant interaction between time and the explicit scores, Pillai's trace = .09, $F(1, 191) = 18.88$, $p < .001$, $\eta_p^2 = .09$. There was also a significant interaction effect between the explicit scores and intervention condition, Pillai's trace = .07, $F(3, 191) = 4.60$, $p = .004$, $\eta_p^2 = .07$. There were no other main or interaction effects (all p's > .069).

To follow up on the interaction between explicit scores and intervention condition a series of Bonferroni corrected one-way ANOVAs (Welch's F) and Games-Howell post hoc tests were conducted (Levene's test of equality of variances p's < .05). There was a significant difference between the intervention groups for the explicit STEM scores at Time 2, Welch's $F(3, 101.05) = 6.84$, $p < .001$, est. $\omega^2 = .08$. In addition to weaker explicit Male-STEM bias at Time 1 (noted above), Games-Howell post hoc tests revealed that the Psychoeducation group's Time 2 STEM scores were 1.95, 95% CI [.40, 3.50] higher than the Control group ($p = .008$), again indicating weaker explicit Male-STEM bias. There was a significant difference for the explicit Arts scores at Time 2, Welch's $F(3, 102.22) = 5.03$, $p = .003$, est. $\omega^2 = .06$. The Psychoeducation group significantly differed from the Control group ($p = .016$) with a difference of -1.27, 95% CI [-2.36, -.18]. This indicated that the Psychoeducation group had weaker explicit Female-Arts bias than the Control group at Time 2 as their scores were closer to the neutral point of 36.

In relation to the interaction between time and explicit scores, STEM scores increased slightly at Time 2 across the whole sample though they still indicated a Male-STEM bias (see Table 1). Arts scores decreased at Time 2 though they still indicated a Female-Arts bias generally. Paired sample t-tests revealed these

differences were statistically significant though effect sizes were small—STEM scores: $t(198) = -3.1$, $p = .002$, Cohen's $d = .22$, 95% CI $[-.63, -.14]$; Arts scores: $t(198) = 3.49$, $p = .001$, Cohen's $d = .25$, 95% CI $[.19, .68]$.

4 Discussion

The primary aim of this study was to assess implicit gender-STEM bias malleability using the IRAP. Secondary aims were to determine which of the three brief interventions were most effective at influencing gender-STEM bias compared to the Control group and whether their effects maintained beyond the initial session. Immediately post-intervention it appeared that implicit gender-STEM bias is indeed malleable in relation to certain contextual factors. Post hoc analysis further highlighted the relations targeted by each intervention. Both the Exemplar and Psychoeducation groups had significantly stronger implicit pro-Women-STEM bias, while only the Psychoeducation group had significantly weaker implicit pro-Men-STEM bias than the Control group. The interventions were less effective for explicit gender-STEM bias. Though explicit scores were not extreme, all groups appeared to demonstrate a Male-STEM/Female-Arts bias at both timepoints. The Psychoeducation group scores, however, were close to the neutral point and significantly differed from the Control group, suggesting less explicit bias.

Levels of bias at Time 2 presented a more complex picture of implicit bias change. The Psychoeducation group and women in the Exemplar group showed stronger pro-Men-STEM bias at Time 2, with medium effect sizes, indicating that their lower pro-Men-STEM bias immediately post-intervention did not maintain. The intervention groups maintained a pro-Women-STEM bias at Time 2. Though there was some reduction in the mean Women-STEM bias for the Exemplar and Psychoeducation groups it was not statistically significant and effect sizes for the differences were small ($d < .5$). This further suggests that interventions may have differential effects on particular relations and these effects may differ in terms of their longevity. The Exemplar condition likely targeted the 'ability stereotype' of gender-STEM bias (Master and Meltzoff 2016) while Psychoeducation likely targeted both 'fit' and 'ability' stereotypes highlighted by Master and Meltzoff (2016). This may contribute to its greater impact on both an implicit and explicit level.

Unexpectedly, at Time 2 Control participants' pro-Women-STEM bias increased significantly, with a medium effect size. This warrants further discussion. Control participants were aware they would be completing the same IRAP and rating scales the following day, so it is highly likely that they discerned that the focus of the study was (in some regard) on gender and STEM. Due to the seemingly unconnected nature of their first task about animals and the gender-STEM IRAP, participants in the Control condition may have ruminated about their experimental experience more than the other groups. Such changes in the experimental context may have influenced the Control participants' responding at Time 2 as implicit responding is sensitive to current (and historical) contextual factors. Gawronski (2017), for example, found that implicit responding has lower temporal stability than explicit responding over time, suggesting the impact of contextual changes (see also Payne et al. 2017).

Additionally, participants may detect which relations they have more difficulty responding quickly to on the IRAP, increasing awareness of their response biases and influencing subsequent responding. While unlikely to create lasting bias change without further reinforcement (see Ebert et al. 2009; see Lai et al. 2016), the impact of completing measures of implicit responding has been explored in the literature. Results have been mixed, sometimes demonstrating effects on explicit attitudes (Menatti et al. 2012), or strengthening positive implicit intervention effects when performance feedback is provided (Pennington et al. 2016). Interestingly, Pennington et al. (2016) found a slight (non-significant) increase in positive attitudes among their control group after they completed an IRAP but only when participants also received IRAP feedback. However, their sample sizes were small ($N = 48$; approx. 12 per group). Completing an IAT can also influence attitudes through associative or analogical learning (e.g., Ebert et al. 2009; Hussey and De Houwer 2018). Recently, Hahn and Gawronski (2019) found predicting one's own implicit biases increased acknowledgement and levels of explicit bias and alignment between implicit and explicit biases, but the influence of completing an IAT with or without feedback was inconsistent—sometimes leading to greater implicit and explicit bias alignment but not greater bias acknowledgement. Participants' increased attention towards their initial, spontaneous affective responses while predicting their implicit biases seemed to be crucial. Therefore, it appears that completing measures of implicit responding or acknowledging one's implicit biases may affect people in varied and complex ways. Lai et al. (2016) suggest using multiple measures of implicit bias to assess what changes may be attributable to the measures and task performance as opposed to implicit bias change. Further large-scale research examining the impact of rumination and completing an IRAP remains warranted particularly within the domain of gender-STEM bias.

We may, therefore, question whether the results of the intervention groups at Time 2 reflect a maintenance of intervention effects or instead the impact of rumination and/or knowledge of the IRAP task. There are two important points to highlight here. First, it appears difficult to fake responding on the IRAP without a concrete strategy and/or detailed instructions which participants are unlikely to spontaneously devise (Drake et al. 2016; Hughes et al. 2016; McKenna et al. 2007). Also, if the Time 2 results represented more effortful, conscious responding, it is interesting to note that there was still an explicit Male-STEM/Female-Arts bias among the groups. One might not have expected this should participants have been engaging in socially desirable responding, for example. It, therefore, seems unlikely that strategic responding was behind the results of Time 2.

The weak (at best) impact of the interventions on explicit gender-STEM bias is reflective of previous research (e.g., Lai et al. 2014). It has been suggested that explicit bias reduction may involve different mechanisms of change (Gawronski and Bodenhausen 2006; Lai et al. 2014) and as such requires tailored approaches. Reducing explicit bias may require longer, more intensive intervention to account for the potentially more complex, elaborated relational responding involved. If these approaches could be incorporated with those that influence implicit bias, then a clearer impact on both implicit and explicit bias may be detected post-intervention. It may also be that intervention effects are very brief and so do not influence explicit

bias assessed after implicit bias (Lai et al. 2014). Again, the longevity of intervention effects warrants further investigation as noted.

4.1 Limitations and future research directions

The following limitations may affect the generalizability and replicability of the results. First, our gender comparison was binary, comparing men and women, as the number of non-binary participants ($n=1$) was too low. Most research in this area has focused on male/female comparisons conceptually and analytically. Subsequent studies should more directly recruit individuals who identify as non-binary or transgender to determine if these results generalize across more gender diverse populations.

Second, we aimed to strengthen a positive relation between women and STEM among those with weak or no positive women-STEM relations, thus women working in/studying STEM subjects were excluded as they have previously demonstrated significant women-STEM relations (Farrell and McHugh 2017). Future studies may utilize a stratified sampling method to examine these interventions with women in STEM. Pre- and post-intervention measurement may be beneficial with this group to assess individual bias change more directly. This was not utilized within the current study so as to reduce the possibility of IRAP practice effects and to conceal the study's true focus (at least until after the intervention task was completed).

Third, the gap between the current sessions was relatively short (16 h minimum). It is unclear whether intervention effects would last beyond this brief timeframe. As noted, it is unlikely that one brief exposure will create lasting change in stimuli relations without further reinforcement within the wider context. This has been reflected by extensive research into the reduction of implicit racial preference (Lai et al. 2016). It may be that longer sessions and/or multiple exposures are required to create more lasting effects.

Fourth, the experimenter was female for all participants which may have influenced subsequent implicit responding. For example, Moss-Racusin et al. (2018) found that viewing female scientists may have encouraged positive attitudes towards women in STEM among their control participants. As such, we cannot rule out the possible influence the experimenter's sex may have had, especially on the Women-STEM IRAP trial-type. Future studies should systematically vary the sex of the experimenter to explore this further.

Fifth, while our sample consisted of adults only, it included both students (STEM and non-STEM; 90%) and non-students (employed and unemployed; 10%). As an initial exploration of the malleability of implicit gender-STEM bias using the IRAP, we did not restrict our sample to students only. This was deemed acceptable as implicit gender-STEM bias has previously been detected among both adult students and non-students (Farrell et al. 2015; Farrell and McHugh 2017; Nosek et al. 2009), and adults can influence students' educational engagement, for example (e.g., Rosenthal et al. 2011). However, it may be the case that interventions are differentially effective for students versus non-students. Future research should examine whether employment status (student versus non-student)

moderates the effectiveness of implicit gender-STEM bias interventions. It will also be important to examine the effect of these interventions with other key groups for the recruitment and retention of female STEM students, such as STEM faculty and adolescents.

Sixth, we should note that the current interventions varied across a number of factors (e.g., duration, direct mention of gender bias) and each represented a 'packaged whole' similar to other interventions in this domain (e.g., Jackson et al. 2014). As we did not manipulate one intervention variable at a time, we cannot pinpoint the exact component which makes each intervention more or less effective. However, we were able to determine which approach as a whole had more of an impact in relation to our Control group. We hypothesize that the extent to which these interventions tapped into key gender-STEM stereotypes may be an important factor in their effectiveness. Future research should, however, investigate the validity of this and consider testing the efficacy of individual intervention components on an implicit level (e.g., inclusion/exclusion of strategies to overcome bias in Psychoeducation interventions; Hennes et al. 2018).

Finally, the Control group results suggest that implicit bias change and/or assessment may be more complex than previously thought. It may be that a relation between women and STEM has a weaker relational history (i.e., a positive relation between women and STEM is less often derived) and is more susceptible to current contextual factors than more established relations. When naïve to the study at Time 1, Control participants exhibited the expected pattern of bias. If participants had remained unaware about the measures to be completed at Time 2 would their bias have remained the same? For example, participants could complete a socially sensitive IRAP on day 1 and then complete this IRAP as well as a gender-STEM IRAP on day 2. This could help determine whether experience with the task in close temporal proximity to completing another IRAP results in stereotype-inconsistent responding on both IRAPs on day 2, or whether only the day 1 IRAP is influenced by rumination on the topic between sessions. This may help us untangle what factors played a role in the unexpected pattern of results found among Control participants at Time 2.

We must also consider the possibility that the IRAP's sensitivity to extraneous variables (e.g., choice of contrasting label statement; Hussey et al. 2016) limits the validity of comparisons across conditions and contexts. This is a potential issue of all measures of implicit bias which are likely susceptible to some degree of influence by contextual factors (Hussey et al. 2016). It has been suggested that the IRAP's sensitivity may make replications across contexts complex (Golijani-Moghaddam et al. 2013). Test–retest reliability of the current IRAP was lower than the median IAT test–retest reliability ($r = .56$; Nosek et al. 2007). However, test–retest reliability has rarely been examined in the IRAP literature (Golijani-Moghaddam et al. 2013). An RFT conceptualization of implicit cognition may also alter expectations regarding test–retest reliability (Golijani-Moghaddam et al. 2013). Other measures of implicit bias may be less sensitive to contextual change; however, this brings its own difficulty as these measures may then fail to detect intervention effects (see Lenton et al. 2009). Longitudinal examination of the malleability of implicit gender-STEM bias and implicit bias measurement is still required.

4.1.1 Refining the interventions

There are a number of suggestions to bear in mind for future research using these or similar interventions. Despite the lack of negative contrast category in the Exemplar condition, gender was made salient (e.g., female scientist passed over for awards in favor of her male colleagues). Therefore, we cannot discount the possibility of a self-generated subtle negative contrast which influenced responding. It may be interesting to expose participants to female scientist exemplars without mentioning their professional and gendered struggles to determine if this approach is as effective. However, role models are more effective when they are perceived to be relatable (e.g., Asgari et al. 2012), and can have a negative impact when they are not (e.g., Betz and Sekaquaptewa 2012; Hoyt and Simon 2011). The exemplars' struggles may make them more relatable. Future studies should carefully select relatable exemplars, perhaps by highlighting how their achievements were accomplished by hard work (e.g., Shin et al. 2016).

Perspective-Taking appeared least effective. This may imply that an increase in self-other overlap and/or empathy does not produce significant gender-STEM bias change. However, the focus of participants' perspective-taking narratives may not have increased a positive relation between women and STEM competency (a key component of gender-STEM bias; Master and Meltzoff 2016), as detailed guidance was not provided for the narratives. Perhaps had participants been asked to imagine they were a female scientist achieving a research breakthrough (highlighting competency) or experiencing bias (evoking empathy) this may have better targeted a positive women-STEM relation. The current perspective-taking may be more beneficial for more negatively evaluated social groups, as women tend to be evaluated positively in a general sense (Eagly et al. 1991).

It may be worthwhile to incorporate the current Psychoeducation piece into a workshop or group format to determine if this would increase or decrease effectiveness (see Moss-Racusin et al. 2018; Smith and Postmes 2011). When the socially shared nature of stereotypes is called into question it can undermine them (Smith and Postmes 2011), which could strengthen the impact of a gender-STEM bias psychoeducation intervention. Perhaps this more intensive approach, incorporating group discussions (e.g., Smith and Postmes 2011) and research-based content delivered by competent individuals or 'experts' (e.g., Moss-Racusin et al. 2018), using inclusive, non-confrontational language (e.g., Jackson et al. 2014) represents a promising avenue of future research to influence both implicit and explicit responding.

4.2 Practical implications

The current results have implications for those developing bias interventions as they demonstrated the factors that increased a positive women-STEM relation, without producing an anti-Men-STEM bias in response. With further validation, the current approaches support unique applications which may be of interest to policymakers. The impact of exemplars advocates for an increase in positive

female exemplars within the media and textbooks where they have been notoriously limited and/or stereotypically portrayed (e.g., Steinke 2012; see Carli et al. 2016). The effect of psychoeducation suggests that counter-stereotypical information may be important (e.g., via the media, education), as is raising awareness regarding implicit bias and how to counter it. It may support the use of tailored and empirically tested bias workshops provided they are based upon current research guidance. This would be of interest to gender equality committees and STEM faculty given the interest in bias workshops as a means of improving gender-STEM diversity. With increasing encouragement to tackle gender equality issues (e.g., Health Research Board Gender Policy, n.d.), it is important that interventions are empirically tested and supported (see Moss-Racusin et al. 2014).

5 Conclusion

The current research suggested that implicit gender-STEM bias is malleable, at least in the short term. It shed light on the particular relations targeted by the interventions, as well as the weaker impact on explicit bias which highlights the need for further research into factors that influence both implicit and explicit responding. The Psychoeducation and Exemplar conditions appeared most effective at strengthening an implicit positive Women-STEM relation. Further validation of these interventions should determine their impact on other important outcomes such as self-efficacy to tackle gender inequality (e.g., Carnes et al. 2015) and relevant behavior change. Additionally, our results call for further research into the longevity of intervention effects and the measurement of implicit bias across contexts. For example, the Time 2 results may reflect a maintenance of intervention effects and the influence of contextual change for the Control group or each groups' IRAP performance may have been altered by their task experience at Time 1.

Sustained cultural change will have stronger lasting effects (see Miller et al. 2015), however, it is important to determine the factors that influence gender-STEM bias and would support such cultural change. Psychoeducation and positive exemplars are promising in this regard. Further opportunities to derive counter-stereotypical relations may strengthen these relations and increase their probability of being emitted. However, further systematic examination is required to determine the limits of these interventions' effectiveness and the particular contexts in which they thrive.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Akosah-Twumasi, P., Emeto, T. I., Lindsay, D., Tsey, K., & Malau-Aduli, B. S. (2018). A systematic review of factors that influence youths career choices—the role of culture. *Frontiers in Education*. https://doi.org/10.3389/feduc.2018.00058.

Ansara, Y. G., & Hegarty, P. (2014). Methodologies of misgendering: Recommendations for reducing cisgenderism in psychological research. *Feminism & Psychology, 24*(2), 259–270. https://doi.org/10.1177/0959353514526217.

Asgari, S., Dasgupta, N., & Stout, J. G. (2012). When do counterstereotypic ingroup members inspire versus deflate? The effect of successful professional women on young women's leadership self-concept. *Personality and Social Psychology Bulletin, 38*(3), 370–383. https://doi.org/10.1177/0146167211431968.

Axelson, R. D., Solow, C. M., Ferguson, K. J., & Cohen, M. B. (2010). Assessing implicit gender bias in medical student performance evaluations. *Evaluation and the Health Professions, 33*(3), 365–385. https://doi.org/10.1177/0163278710375097.

Barnes Holmes, D., Barnes Holmes, Y., Smeets, P. M., Cullinan, V., & Leader, G. (2004). Relational frame theory and stimulus equivalence: Conceptual and procedural issues. *International Journal of Psychology and Psychological Therapy*, *4*(2), 181–214. https://search.proquest.com/openview/74a3b71e11f627c02a6320b1942036bd/1?pq-origsite=gscholar&cbl=29514.

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist, 32*(7), 169–177. Retrieved March 3rd 2019 from https://www.researchgate.net/profile/Ian_Stewart3/publication/241832168_Do_You_Really_Know_What_You_Believe_Developing_the_Implicit_Relational_Assessment_Procedure_IRAP_as_a_Direct_Measure_of_Implicit_Beliefs/links/5613809e08aea9fb51c303ce/Do-You-Really-Know-What-You-Believe-Developing-the-Implicit-Relational-Assessment-Procedure-IRAP-as-a-Direct-Measure-of-Implicit-Beliefs.pdf.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record, 60*(3), 527. https://doi.org/10.1007/BF03395726.

Betz, D. E., & Sekaquaptewa, D. (2012). My fair physicist? Feminine math and science role models demotivate young girls. *Social Psychological and Personality Science, 3*(6), 738–746. https://doi.org/10.1177/1948550612440735.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81*(5), 828. https://doi.org/10.1037//0022-3514.81.5.828.

Carli, L. L., Alawa, L., Lee, Y., Zhao, B., & Kim, E. (2016). Stereotypes about gender and science: Women ≠ Scientists. *Psychology of Women Quarterly, 40*(2), 244–260. https://doi.org/10.1177/0361684315622645.

Carnes, M., Bartels, C. M., Kaatz, A., & Kolehmainen, C. (2015). Why is John more likely to become department chair than Jennifer? *Transactions of the American Clinical and Climatological Association, 126*, 197. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4530686/.

Carnes, M., Devine, P. G., Isaac, C., Manwell, L. B., Ford, C. E., Byars-Winston, A., et al. (2012). Promoting institutional change through bias literacy. *Journal of Diversity in Higher Education, 5*(2), 63–77. https://doi.org/10.1037/a0028128.

Carnes, M., Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., et al. (2015b). Effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. *Academic Medicine: Journal of the Association of American Medical Colleges, 90*(2), 221. https://doi.org/10.1097/ACM.0000000000000552.

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology, 97*(6), 1045. https://doi.org/10.1037/a0016239.

Cullen, C., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The Implicit Relational Assessment Procedure (IRAP) and the malleability of ageist attitudes. *The Psychological Record, 59*(4), 591. https://doi.org/10.1007/bf03395683.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on automatic gender stereotyping. *Journal of Experimental Social Psychology, 40,* 642–658. https://doi.org/10.1016/j.jesp.2004.02.003.

Dasgupta, N. G., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81,* 800–814. https://doi.org/10.1037//0022-3514.81.5.800.

De Houwer, J., & De Bruycker, E. (2007). The implicit association test outperforms the extrinsic affective Simon task as an implicit measure of inter-individual differences in attitudes. *British Journal of Social Psychology, 46*(2), 401–421. https://doi.org/10.1348/014466606x130346.

Dovidio, J. F., Ten Vergert, M., Stewart, T. L., Gaertner, S. L., Johnson, J. D., Esses, V. M., et al. (2004). Perspective and prejudice: Antecedents and mediating mechanisms. *Personality and Social Psychology Bulletin, 30*(12), 1537–1549. https://doi.org/10.1177/0146167204271177.

Drake, C. E., Seymour, K. H., & Habib, R. (2016). Testing the IRAP: exploring the reliability and fakability of an idiographic approach to interpersonal attitudes. *The Psychological Record, 66*(1), 153–163. https://doi.org/10.1007/s40732-015-0160-1.

Eagly, A. H., Mladinic, A., & Otto, S. (1991). Are women evaluated more favorably than men? An analysis of attitudes, beliefs, and emotions. *Psychology of Women Quarterly, 15*(2), 203–216. https://doi.org/10.1111/j.1471-6402.1991.tb00792.x.

Ebert, I. D., Steffens, M. C., Von Stülpnagel, R., & Jelenec, P. (2009). How to like yourself better, or chocolate less: Changing implicit attitudes with one IAT task. *Journal of Experimental Social Psychology, 45*(5), 1098–1104. https://doi.org/10.1016/j.jesp.2009.06.008.

Edwards, D. J., McEnteggart, C., Barnes-Holmes, Y., Lowe, R., Evans, N., & Vilardaga, R. (2017). The impact of mindfulness and perspective-taking on implicit associations toward the elderly: a relational frame theory account. *Mindfulness*. https://doi.org/10.1007/s12671-017-0734-x.

Ehrlinger, J., Plant, E. A., Hartwig, M. K., Vossen, J. J., Columb, C. J., & Brewer, L. E. (2017). Do gender differences in perceived prototypical computer scientists and engineers contribute to gender gaps in computer science and engineering? *Sex Roles*. https://doi.org/10.1007/s11199-017-0763-x.

Farrell, L., Cochrane, A., & McHugh, L. (2015). Exploring attitudes towards gender and science: the advantages of an IRAP approach versus the IAT. *Journal of Contextual Behavioral Science, 4*(2), 121–128. https://doi.org/10.1016/j.jcbs.2015.04.002.

Farrell, L., & McHugh, L. (2017). Examining gender-STEM bias among STEM and non-STEM students using the Implicit Relational Assessment Procedure (IRAP). *Journal of Contextual Behavioral Science, 6*(1), 80–90. https://doi.org/10.1016/j.jcbs.2017.02.001.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*(4), 708. https://doi.org/10.1037//0022-3514.78.4.708.

Gatta, M., & Trigg, M. (2001). *Bridging the gap: Gender equity in science, engineering and technology.* Rutgers University, Center for Women and Work. Retrieved 13th February 2014 from http://opas.ous.edu/Committees/Resources/Research_papers/Gatta_Trigg_Bridging.pdf.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692. https://doi.org/10.1037/0033-2909.132.5.692.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: a longitudinal analysis. *Personality and Social Psychology Bulletin, 43*(3), 300–312. https://doi.org/10.1177/0146167216684131.

Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The implicit relational assessment procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science, 2*(3), 105–119. https://doi.org/10.1016/j.jcbs.2013.05.002.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education, 74*(3), 249–266. https://doi.org/10.3200/jexe.74.3.249-266.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464. https://doi.org/10.1037//0022-3514.74.6.1464.

Hahn, A., & Gawronski, B. (2019). Facing one's implicit biases: From awareness to acknowledgment. *Journal of Personality and Social Psychology, 116*(5), 769. https://doi.org/10.1037/pspi0000155.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: a post-skinnerian account of human language and cognition.* New York, NY: Plenum. https://books.google.ie/books?id=89iIBwAAQBAJ&lr=&source=gbs_navlinks_s.

Hennes, E. P., Pietri, E. S., Moss-Racusin, C. A., Mason, K. A., Dovidio, J. F., Brescoll, V. L., et al. (2018). Increasing the perceived malleability of gender bias using a modified Video Intervention for Diversity in STEM (VIDS). *Group Processes & Intergroup Relations, 21*(5), 788–809. https://doi.org/10.1177/1368430218755923.

Hinton, P. R., Brownlow, C., McMurray, I., & Cozens, B. (2004). SPSS explained. Routledge. https://doi.org/10.4324/9780203642597.

Hooper, N., Villatte, M., Neofotistou, E., & McHugh, L. (2010). The effects of mindfulness versus thought suppression on implicit and explicit measures of experiential avoidance. *International Journal of Behavioral Consultation and Therapy, 6*(3), 233. https://doi.org/10.1037/h0100910.

Hoyt, C. L., & Simon, S. (2011). Female leaders: Injurious or inspiring role models for women? *Psychology of Women Quarterly, 35*(1), 143–157. https://doi.org/10.1177/0361684310385216.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (Vol. 498). Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/047178947X.

Hughes, S., Barnes-Holmes, D., & Smyth, S. (2017). Implicit cross-community biases revisited: Evidence for ingroup favoritism in the absence of outgroup derogation in Northern Ireland. *The Psychological Record, 67*(1), 97–107. https://doi.org/10.1007/s40732-016-0210-3.

Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: a voyage through implicit cognition research. *Journal of Contextual Behavioral Science, 1*(1–2), 17–38. https://doi.org/10.1016/j.jcbs.2012.09.003.

Hughes, S., Hussey, I., Corrigan, B., Jolie, K., Murphy, C., & Barnes-Holmes, D. (2016). Faking revisited: Exerting strategic control over performance on the Implicit Relational Assessment Procedure. *European Journal of Social Psychology, 46*(5), 632–648. https://doi.org/10.1002/ejsp.2207.

Hussey, I., & De Houwer, J. (2018). Implicit association test as an analogical learning task. *Experimental Psychology, 65*(5), 272–285. https://doi.org/10.1027/1618-3169/a000416.

Hussey, I., Mhaoileoin, D. N., Barnes-Holmes, D., Ohtsuki, T., Kishita, N., Hughes, S., et al. (2016). The IRAP is nonrelative but not acontextual: changes to the contrast category influence men's dehumanization of women. *The Psychological Record, 66*(2), 291–299. https://doi.org/10.1007/s40732-016-0171-6.

Isaac, C., Chertoff, J., Lee, B., & Carnes, M. (2011). Do students' and authors' genders affect evaluations? A linguistic analysis of medical student performance evaluations. *Academic Medicine, 86*(1), 59. https://doi.org/10.1097/acm.0b013e318200561d.

Jackson, S. M., Hillard, A. L., & Schneider, T. R. (2014). Using implicit bias training to improve attitudes toward women in STEM. *Social Psychology of Education, 17*(3), 419–438. https://doi.org/10.1007/s11218-014-9259-5.

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*(3), 137–146. https://doi.org/10.1027/1864-9335/a000020.

Kaatz, A., & Carnes, M. (2014). Stuck in the out-group: Jennifer can't grow up, Jane's invisible, and Janet's over the hill. *Journal of Women's Health, 23*(6), 481–484. https://doi.org/10.1089/jwh.2014.4766.

Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes, gender identification, and math-related outcomes: a prospective study of female college students. *Psychological Science, 18*(1), 13–18. https://doi.org/10.1111/j.1467-9280.2007.01841.x.

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication, 35*(5), 603–625. https://doi.org/10.1177/1075547012472684.

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7*(5), 315–330. https://doi.org/10.1111/spc3.12023.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E. L., Joy-Gaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*(4), 1765–1785. https://doi.org/10.1037/a0036260.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*(8), 1001–1016. https://doi.org/10.1037/xge0000179.

Lenton, A. P., Bruder, M., & Sedikides, C. (2009). A meta-analysis on the malleability of automatic gender stereotypes. *Psychology of Women Quarterly, 33,* 183–196. https://doi.org/10.1111/j.1471-6402.2009.01488.x.

Leslie, S. J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347*(6219), 262–265. https://doi.org/10.1126/science.1261375.

Margolis, J., Fisher, A., & Miller, F. (2000). The anatomy of interest: Women in undergraduate computer science. *Women's Studies Quarterly, 1*(2), 104–127. https://www.jstor.org/stable/40004448.

Master, A., Cheryan, S., & Meltzoff, A. N. (2016). Computing whether she belongs: Stereotypes undermine girls' interest and sense of belonging in computer science. *Journal of Educational Psychology, 108*(3), 424–437. https://doi.org/10.1037/edu0000061.

Master, A., & Meltzoff, A. N. (2016). Building bridges between psychological science and education: Cultural stereotypes, STEM, and equity. *Prospects, 46*(2), 215–234. https://doi.org/10.1007/s11125-017-9391-z.

McKenna, I. M., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2007). Testing the Fake-ability of the Implicit Relational Assessment Procedure (IRAP): the first study. *International Journal of Psychology and Psychological Therapy, 7*(2), 253–268. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.5392&rep=rep1&type=pdf.

McKenna, I., Hughes, S., Barnes-Holmes, D., De Schryver, M., Yoder, R., & O'Shea, D. (2016). Obesity, food restriction, and implicit attitudes to healthy and unhealthy foods: Lessons learned from the implicit relational assessment procedure. *Appetite, 100,* 41–54. https://doi.org/10.1016/j.appet.2016.02.013.

Menatti, A. R., Smyth, F. L., Teachman, B. A., & Nosek, B. A. (2012). Reducing stigma toward individuals with mental illnesses: a brief, online intervention. *Stigma Research and Action, 1,* 9–21. https://doi.org/10.31234/osf.io/ghfa3.

Miller, D. I., Eagly, A. H., & Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology, 107*(3), 631–644. https://doi.org/10.1037/edu0000005.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109.

Moss-Racusin, C. A., Pietri, E. S., Hennes, E. P., Dovidio, J. F., Brescoll, V. L., Roussos, G., et al. (2018). Reducing STEM gender bias with VIDS (video interventions for diversity in STEM). *Journal of Experimental Psychology: Applied, 24*(2), 236–260. https://doi.org/10.1037/xap0000144.

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2014). Scientific diversity interventions. *Science, 343*(6171), 615–616. https://doi.org/10.1126/science.1245936.

Moss-Racusin, C. A., van der Toorn, J., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2016). A "scientific diversity" intervention to reduce gender bias in a sample of life scientists. *CBE-Life Sciences Education, 15*(3), ar29. https://doi.org/10.1187/cbe.15-09-0187.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044–1045. https://doi.org/10.1093/beheco/arh107.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine of the National Academies. (2006). *Beyond biases and barriers: Fulfilling the potential of women in academic science and engineering*. Washington, DC: National Academies Press. https://doi.org/10.17226/11741.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The implicit association test at age 7: a methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). London: Psychology Press. https://faculty.washington.edu/agg/pdf/Nosek%20&%20al.IATatage7.2007.pdf.

Nosek, B. A., & Smyth, F. L. (2011). Implicit social cognitions predict sex differences in math engagement and achievement. *American Educational Research Journal, 48*(5), 1125–1156. https://doi.org/10.3102/0002831211410683.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., et al. (2009). National differences in gender—science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248. https://doi.org/10.1080/1047840x.2017.1335568.

Pennington, C. R., Campbell, C., Monk, R. L., & Heim, D. (2016). The malleability of stigmatizing attitudes: Combining imagined social contact with implicit attitude feedback. *American Journal of Psychiatric Rehabilitation, 19*(3), 175–195. https://doi.org/10.1080/15487768.2016.1171175.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ (Clinical Research ed.), 316*(7139), 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236.

Qualtrics. (2017). Provo, Utah, USA. Available from https://www.qualtrics.com.

Ramsey, L. R., Betz, D. E., & Sekaquaptewa, D. (2013). The effects of an academic environment intervention on science identification among women in STEM. *Social Psychology of Education, 16*(3), 377–397. https://doi.org/10.1007/s11218-013-9218-6.

Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M. A., & De Raedt, R. (2013). Self-esteem revisited: Performance on the implicit relational assessment procedure as a measure of self-versus ideal self-related cognitions in dysphoria. *Cognition and Emotion, 27*(8), 1441–1449. https://doi.org/10.1080/02699931.2013.786681.

Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences, 111*(12), 4403–4408. https://doi.org/10.1073/pnas.1314788111.

Rosenthal, L., London, B., Levy, S. R., & Lobel, M. (2011). The roles of perceived identity compatibility and social support for women in a single-sex STEM program at a co-educational University. *Sex Roles, 65*(9–10), 725–736. https://doi.org/10.1007/s11199-011-9945-0.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology, 1*(1), 43–46. https://doi.org/10.1097/00001648-199001000-00010.

Rudman, L. A., Ashmore, R. D., & Gary, M. L. (2001). "Unlearning" automatic biases: the malleability of implicit stereotypes and prejudice. *Journal of Personality and Social Psychology, 81*, 856–868. https://doi.org/10.1037/0022-3514.81.5.856.

Sánchez de Madariaga, I., Raudma, T., Eichenberger, T., Hogan, A., Pollitzer, E., Rees, T., et al. (2012). *Structural change of research institutions. Enhancing excellence, gender equality and efficiency in research and innovation. Report of the expert group on structural change.* Brussels, European Commission. Retrieved 5th May 2017 from https://ec.europa.eu/research/science-society/document_library/pdf_06/structural-changes-final-report_en.pdf.

Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: the role of stereotype endorsement in women's experience in the math domain. *Sex Roles, 50*(11–12), 835–850. https://doi.org/10.1023/b:sers.0000029101.74557.a0.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles, 57*(7–8), 509–514. https://doi.org/10.1007/s11199-007-9291-4.

Schwab, K., Samans, R., Zahidi, S., Leopold, T. A., Ratcheva, V., Hausmann, R., & D'Andrea Tyson, L. (2016). Global gender gap report 2016. In *World economic forum*. Retrieved 26th April 2017 from http://www3.weforum.org/docs/GGGR16/WEF_Global_Gender_Gap_Report_2016.pdf.

Shih, M., Wang, E., Trahan Bucher, A., & Stotzer, R. (2009). Perspective taking: Reducing prejudice towards general outgroups and specific individuals. *Group Processes & Intergroup Relations, 12*(5), 565–577. https://doi.org/10.1177/1368430209337463.

Shin, J. E. L., Levy, S. R., & London, B. (2016). Effects of role model exposure on STEM and non-STEM student engagement. *Journal of Applied Social Psychology, 46*(7), 410–427. https://doi.org/10.1111/jasp.12371.

Smeding, A. (2012). Women in science, technology, engineering, and mathematics (STEM): an investigation of their implicit gender stereotypes and stereotypes' connectedness to Math performance. *Sex Roles, 67*(11–12), 617–629. https://doi.org/10.1007/s11199-012-0209-4.

Smith, L. G., & Postmes, T. (2011). Shaping stereotypical behaviour through the discussion of social stereotypes. *British Journal of Social Psychology, 50*(1), 74–98. https://doi.org/10.1348/014466610x 500340.

Smyth, F. L., & Nosek, B. A. (2015). On the gender–science stereotypes held by scientists: Explicit accord with gender-ratios, implicit accord with scientific identity. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.00415.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's Math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373.

Steinke, J. (2012). Portrayals of female scientists in the mass media: End times for a media history paradigm. *The International Encyclopedia of Media Studies*. https://doi.org/10.1002/9781444361506.wbiems070.

Stout, J. G., Dasgupta, N., Hunsinger, M., & McManus, M. A. (2011). STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology, 100*(2), 255–270. https://doi.org/10.1037/a0021385.

Summers, L. H. (2005). *Remarks at NBER conference on diversifying the science and engineering workforce,* January 2005. Transcript taken from http://www.president.harvard.edu/speeches/2005/nber.html.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*, 5th edn. Allyn & Bacon/Pearson Education. https://dl.acm.org/citation.cfm?id=1213888.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*, 6th edn. Pearson Education. https://www.pearson.com/us/higher-education/program/Tabachnick-Using-Multivariate-Statistics-6th-Edition/PGM332849.html.

Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology, 100*(6), 1027–1042. https://doi.org/10.1037/a0022308.

Todd, A. R., Galinsky, A. D., & Bodenhausen, G. V. (2012). Perspective taking undermines stereotype maintenance processes: Evidence from social memory, behavior explanation, and information solicitation. *Social Cognition, 30*(1), 94–108. https://doi.org/10.1521/soco.2012.30.1.94.

UNESCO. (2017). *Women in science*. Retrieved 26th April 2017 from http://www.uis.unesco.org/_LAYOUTS/UNESCO/women-in-science/index.html#overview!lang=en.

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry, 48,* 59–65. https://doi.org/10.1016/j.jbtep.2015.01.004.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences, 112*(17), 5360–5365. https://doi.org/10.1073/pnas.1418878112.

Yee, N., & Bailenson, J. N. (2006). Walk a mile in digital shoes: the impact of embodied perspective-taking on the reduction of negative stereotyping in immersive virtual environments. *Proceedings of PRESENCE, 24*, 147–156. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.545.4279&rep=rep1&type=pdf.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*(1), 173–181. https://doi.org/10.1348/000711004849222.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr Lynn Farrell is a Research Fellow at the School of Psychology, Queen's University Belfast. Her research interests include the nature and malleability of gender bias and stereotypes, particularly in relation to STEM fields, examining the effectiveness of Gender Equality Initiatives, and the application of a Relational Frame Theory perspective to social psychological phenomena.

Dr Finiki Nearchou is an Assistant Professor with the School of Psychology, University College Dublin (UCD) and a UCD Ad Astra Fellow. Her research interests include identifying predictors of various domains of health and well-being in children, adolescents and young adults, stigma, and psychometric tool development and validation.

Prof Louise McHugh is a Professor at the School of Psychology, UCD. Her research interests are centered on the experimental analysis of language and cognition from a behavior analytic and Relational Frame Theory perspective, including especially the development of complex cognitive skills such as perspective-taking, and the process-level investigation of behavioral and cognitive psychotherapies including Acceptance and Commitment Therapy.

Affiliations

Lynn Farrell[1,2] · **Finiki Nearchou[1]** · **Louise McHugh[1]**

Finiki Nearchou
niki.nearchou@ucd.ie

Louise McHugh
louise.mchugh@ucd.ie

[1]   School of Psychology, University College Dublin, Belfield, Dublin 4, Ireland

[2]   School of Psychology, Queen's University Belfast, University Road, Belfast, Northern Ireland, UK